МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ УКРАИНЫ

ХАРЬКОВСЬКИЙ НАЦИОНАЛЬНЫЙ УНИВЕРСИТЕТ ГОРОДСКОГО ХОЗЯЙСТВА имени А. Н. БЕКЕТОВА

А. В. Белогурова

ТЕОРИЯ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

ЧАСТЬ 2. МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

КОНСПЕКТ ЛЕКЦИЙ

(для студентов 2-го курса дневной и заочной форм обучения, направления подготовки 6.030504 – Экономика предприятия, 6.030509 – Учет и аудит)

Харьков – ХНУГХ – 2015

Белогурова А. В. Теория вероятностей и математическая статистика. Часть 2. Математическая статистика: Конспект лекций (для студентов 2-го курса дневной и заочной форм обучения, направления подготовки 6.030504 – Экономика предприятия, 6.030509 – Учет и аудит) / А. В. Белогурова; Харьков. нац. ун-т гор. хоз-ва им. А. Н. Бекетова. – Харьков : ХНУГХ, 2015. – 26 с.

Автор: к.т.н., доц. А.В. Белогурова

Конспект лекций построен в соответствии с требованиями кредитно-модульной системы организации учебного процесса и согласован с ориентировочной структурой содержания учебной дисциплины, рекомендованной Европейской Кредитно-Трансферной Системой (ECTS).

Рекомендовано для студентов специальностей «Экономика предприятий», «Учет и аудит».

Рецензент: доц., к. ф.-м. н. А. Б. Костенко

Утверждено на заседании кафедры Прикладной математики и информационных технологий, протокол № 1 от 30 августа 2011 г.

ОГЛАВЛЕНИЕ

СМ2. МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

1 Основные определения математической статистики

Математическая статистика занимается *установлением закономерностей*, которым подчинены массовые случайные явления. Установление закономерностей основано на изучении *статистических данных* – результатах наблюдений, которые можно рассматривать как реализации некоторой случайной величины X. Наблюдаемые значения этой величины $x_1, x_2, \ldots, x_n$ (перечень значений признака X) называют *случайной выборкой*, а количество наблюдений n – *объемом выборки*. Как случайная величина X, так и ее значения могут быть векторами. Множество возможных значений, которые могут наблюдаться при реализации эксперимента, образуют *выборочное пространство* (или *генеральную совокупность*). Например, наблюдаемый признак X – часовая производительность (количество единиц изделий, обработанных за час) одного рабочего в течение календарного периода.

Задачи математической статистики возникают, когда *функция распределения* наблюдаемого признака X неизвестна, при этом методы статистического анализа позволяют получить информацию о различных закономерностях в генеральной совокупности. Задачи математической статистики можно разделить на две большие группы: *первая группа задач* – определение способов сбора и группировки статистических сведений, а *вторая* – статистический анализа статистических данных.

Статистический анализ, как правило, начинается с предварительного анализа и обработки статистических данных. На этом этапе необходимо четко определить *цели анализа*, получить и первично *обработать данные*, определить их *тип и структуру*, подобрать и обосновать *статистические методы*, с помощью которых можно достичь целей анализа, подготовить данные для применения выбранных статистических методов и только после этого выполнить статистический анализ данных. Например, если возникли подозрения, что выборка имеет значения, которые являются случайными или резко выделяются на фоне остальных выборочных значений (*выбросы*), то следует провести *цензуирование выборки* (удаления выбросов из выборки).

К статистическому анализу можно отнести следующие задачи: установление выборочного распределения, определение статистических характеристик выборки, проверка всевозможных статистических гипотез,

сравнение нескольких одномерных выборок (одинаковое распределение, одинаковые выборочные параметры), корреляционный анализ, сравнение зависимых выборок (дисперсионный анализ), регрессионный анализ.

1.1 Способы отбора данных

Иногда проводят *сплошное обследование* (обследуется каждый объект из *генеральной совокупности*), но на практике оно применяется редко, из-за большого количества объектов, или из-за невозможности сплошного обследования, или обследование объекта связано с его уничтожением, или требует больших материальных затрат. Поэтому чаще проводят *выборочное обследование*.

Выборка может быть *повторной* и *бесповторной*. *Повторной* называют выборку, при которой объект (перед отбором следующего) возвращается в генеральную совокупность. *Бесповторной* называют выборку, при которой отобранный объект в генеральную совокупность не возвращается.

Выборка *репрезентативна* (представительна), если *вероятностные свойства выборки* совпадают с *вероятностные свойства генеральной совокупности*. Представительную выборку можно получить, если выбрать объекты для исследования случайно, т.е. все объекты генеральной совокупности с одинаковой вероятностью могут подвергнуться исследованию.

На практике применяют следующие *способы отбора данных*:

1. *Отбор, не требующий расчленения генеральной совокупности* (например, с помощью таблицы «случайных чисел» отбираются объекты, номера которых совпадают с числами из таблицы):
   - *Простой случайный бесповторный отбор* (повторяющиеся числа таблицы «случайных чисел» пропускаются);
   - *Простой случайный повторный отбор*.
2. *Генеральная совокупность расчленяется на части*:
   - *Типический отбор*, например отбор деталей производится не из всех деталей, а из деталей каждого станка. Типический отбор целесообразен, если, например, среди машин, изготавливающих детали, есть более и менее изношенные;
   - *Механический отбор* – генеральная совокупность «механически» делится на такое количество групп, сколько вариантов должно войти в выборку, и из каждой группы берут один объект. Например, если нужно отобрать 5% деталей, то отбирают каждую двадцатую деталь;

- *Серийный отбор* – системному обследованию производится серии целиком. Например, если изделия изготавливаются большой группой станков-автоматов, то подвергают сплошному обследованию продукцию только нескольких станков. Серийным отбором пользуются тогда, когда обследуемый признак колеблется в различных сериях незначительно.

1.2 Предварительный анализ выборочных данных

Предварительный анализ выборочных данных заключается в подготовке данных к последующему анализу. Подготовка данных может включать различные действия, определяемые целями статистического анализа. Например, провести *цензуирование выборки*, если выборка имеет выбросы. Если необходима интервальная оценка неизвестных параметров распределения, то предварительным этапом можно считать проверку гипотезы о нормальности выборочного распределения. Чаще всего целью статистического анализа может быть определение типа выборочного распределения, а на предварительном этапе строят гистограммы выборочного распределения и подсчитывают различные статистические характеристики выборки.

1.2.1 Цензуирование выборки – это процесс удаления из выборки выбросов. В зависимости от природы выбросов (это ошибки наблюдений или артефакты, привнесенные человеком, либо корректные, но «отличающиеся от остальных» значения данных) проблему выбросов решают по-разному. Если это элементарная ошибка наблюдений, то значение по возможности нужно просто откорректировать. Если это артефакт, не подлежащий корректировке, то его удаляют. Если есть убедительные подтверждения тому, что значения-выбросы не принадлежат генеральной совокупности, то их также удаляют.

Существует несколько основных подходов к идентификации выбросов, которые можно разделить на две группы: методы, основанные на априорной вероятности о распределении генеральной совокупности, и непараметрические методы, не использующие информацию о распределении генеральной совокупности.

1.2.2 Построение гистограмм, полигонов и эмпирических функций распределения может также считаться предварительным анализом данных. Рассмотрим это подробнее.

Совокупность всех значений какого-то признака объектов, например время работы ламп накаливания, *называется* *генеральной совокупностью*. *Выборочной совокупностью* или просто *выборкой* называют совокупность объектов случайно отобранных из генеральной совокупности. При исследовании объектов можно фиксировать значение *одного или нескольких признаков*, т.е. выборка может быть *одномерной*, *двумерной*, *трехмерной* и т.д. *Объемом совокупности* (*выборочной* или *генеральной*) называют число объектов этой совокупности (*n* или *N* соответственно).

Пусть из генеральной совокупности извлечена выборка, причем *варианта* x_1 наблюдалась m_1 раз, x_2 – m_2 раз, …, x_K – m_K раз и $\sum m_i = n$ (сумма частот равна объему выборки). Последовательность *вариант* x_i (записанных в возрастающем порядке) и их *частот* m_i называется *рядом распределения* (*эмпирическим рядом распределения*). Отношения $\frac{m_i}{n} = P_i^*$ называются *относительными частотами* или *эмпирической вероятностью*. Ряды распределения, образованные из значений случайной величины, характеризующей *качественный признак* (отличный, хороший, посредственный), называют *атрибутивными рядами*. А ряды распределения, образованные из значений случайной величины, характеризующей *количественный признак* события или явления, называют *вариационными рядами.*

Варианты признака *X*	Частоты	Эмпир. Вер-сть
x_1	m_1	p_1^*
x_2	m_2	p_2^*
…	…	…
x_k	m_k	p_k^*
Объем выборки	$n = \sum_{i=1}^{k} m_i$	1

Построение эмпирических графиков и диаграмм позволяет установить на первом этапе исследования, к какому типу теоретических распределений ближе всего полученное эмпирическое распределение, что облегчает выбор конкретных технических приемов обработки исходных данных.

Эмпирические ряды распределения, получаемые при обработке первичных статистических данных, оформляются *в таблицах* (рис. 1.) или *изображаются графически посредством геометрических образов* – точек, линий и фигур в различных сочетаниях. Здесь x_i – наблюдаемое значение признака *X*, а m_i – частота, с которой встречается x_i. *Относительной частотой* (*эмпирической вероятностью*) p_i^* называется отношение частоты m_i к объему выборки n.

Вариационный ряд можно изобразить графически в виде *полигона* или *гистограммы*. *Полигоном частот* называют ломаную, соединяющую точки с координатами (x_1, m_1), (x_2, m_2), …, (x_k, m_k). *Полигоном относительных частот* называют ломанную, отрезки которой соединяют точки с координатами (x_1, p_1^*), (x_2, p_2^*), …, (x_k, p_k^*) (рис. 2.).

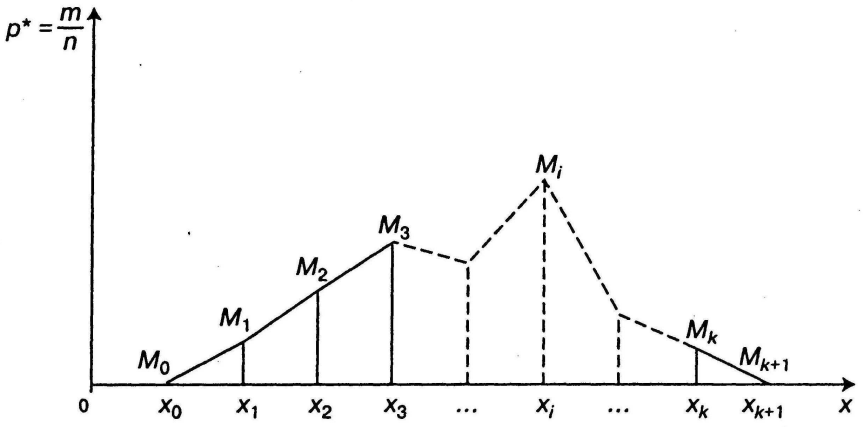


Рисунок 2 – Полигон относительных частот

Гистограмма распределения реализаций случайной величины применяется для графического изображения интервальных рядов распределения. Она представляет собой многоугольник, построенный с помощью смежных прямоугольников. В случае непрерывных равных интервалов с шириной интервала $l = \Delta x$ гистограмма выглядит следующим образом (рис. 3).

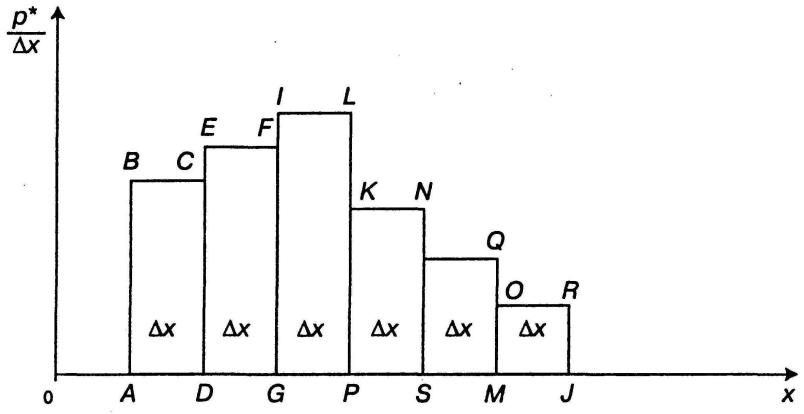


Рисунок 3 – Гистограмма распределения

Вариационный ряд по интервалам чаще изображают с помощью гистограммы частот.

Гистограммой частот называют ступенчатую фигуру, состоящую из прямоугольников, основаниями которых служат частичные интервалы l, а высоты равны отношению $h=\frac{m_i}{l}$ (плотность частоты). Площадь гистограммы частот равна сумме частот, т.е. объему выборки.

Гистограммой относительных частот называют ступенчатую фигуру, состоящую из прямоугольников, основаниями которых служат частичные интервалы l, а высоты равны отношению $h=\frac{p_i^*}{l}$ (плотность относительной частоты). Площадь гистограммы относительных частот равна сумме всех относительных частот, т.е. единице.

Эмпирической функцией распределения (*функцией распределения выборки*) называют функцию $F^*(x)$, определяющую для каждого значения x относительную частоту события $X<x$. $F^*(x)=\frac{n_x}{n}$, где $n_x=\sum_{i=1}^{x} n_i$ – число вариант, меньших x, n – объем выборки. Часто к обычному вариационному ряду добавляют относительные частоты и значения эмпирической функции распределения (табл. 1.)

Таблица 1 – Вариационный ряд

x_i	m_i	P_i *	$F^*(x)$
x_1	m_1	$p_1=\frac{m_1}{n}$	p_1
x_2	m_2	$p_2=\frac{m_2}{n}$	p_1+p_2
…	…	...	...
x_k	m_k	$p_k=\frac{m_k}{n}$	$\sum_{i=1}^{k} p_i=1$
Объем выборки	$n=\sum_{i=1}^{k} m_i$	$\sum_{i=1}^{k} p_i=1$	–

В отличии от эмпирической функции распределения выборки, интегральную *функцию* $F(x)$ *распределения генеральной совокупности* называют *теоретической функцией распределения*. Различие между эмпирической и теоретической функциями состоит в том, что теоретическая функция $F(x)$ определяет *вероятность события* $X<x$, а эмпирическая функция $F^*(x)$ определяет *относительную частоту* этого же события. $F^*(x)$ – является оценкой ординаты интегральной функции теоретического распределения.

Пример. Задано распределение частот выборки объема 20:

x_i 2 6 12

m_i 3 10 7

Определить распределение относительных частот и эмпирическую функцию распределения.

x_i	m_i	P_i *	$F^*(x)$
2	*3*	0.15	0.15
6	*10*	0.50	0.65
12	7	0.35	1
Объем выборки	$n = 20$	$\sum_{i=1}^{k} p_i = 1$	–

Покажем, каким образом производится обработка статистического материала, если производится анализ некоторого непрерывного признака *X*. Весь диапазон значений непрерывной случайной величины *X* разбивается на *K* интервалов (карманов), длинной *l*. Далее подсчитывается количество значений случайной величины *X*, приходящейся на каждый *i*-ый интервал (m_i частота попадания в *i*-ый интервал), и определяется ее эмпирическая вероятность $P_i^* = \frac{m_i}{n}$. Если случайная величина *X* принимает значение, попадающее на границу *i*-го и (*i* + 1) –го интервалов, то это значение учитывается в числе попаданий в (*i* + 1) – й интервал. Таким образом, получим вариационный ряд по интервалам.

Количество карманов может быть различным, например его можно определить по таблице 2.

Таблица 2 – Зависимость количества интервалов от объема выборки

Объем выборки	100	200	400	600	800	1000	1500	2000
Кол-во интервалов	12	16	20	24	27	30	35	37

Для этого можно использовать формулу Стерджесса $K = [1 + 3{,}22 \cdot \ln(n)]$, где квадратные скобочки обозначают целую часть числа. Приведу еще несколько формул:

$K = 10 \cdot \lg(n)$, при этом $K \in [5{,}30]$;

$K = 5 \cdot \lg(n)$ и $K \in [6{,}20]$;

$K = [3{,}26 \cdot \lg(n) + 0{,}5] + 1$, если $n \le 100$ и $K = \min([0{,}1n], 25) + 1$, $n > 100$;

$K = \left[4 \cdot \left(0{,}75(n-1)^2\right)^{1/5} \right]$, если $n > 200$ и $K = [0{,}2n]$, если $n \le 200$;

$K = \min([\sqrt{n}], 30)$.

Какие бы формулы не использовались для вычисления K, следует помнить, что при слишком большом значении K вид распределения искажается случайными значениями частот (поскольку интервалы очень короткие). А при малом числе интервалов сглаживаются и нивелируются характерные особенности распределения (например, наличие двух близкорасположенных мод). Поэтому для качественного анализа строят гистограммы при нескольких значениях K. Если количество карманов известно, то длина интервала определяется формулой $l = \frac{x_{\max} - x_{\min}}{K}$.

Можно пойти другим путем, сначала найти длину интервала, например по формуле: $l = \frac{x_{\max} - x_{\min}}{1 + 3.21 \cdot \lg n}$, где $x_{\max} - x_{\min}$ − размах вариации случайной величины X, тогда число интервалов будет равно $K = \frac{x_{\max} - x_{\min}}{l}$. Если K нецелое, то его округляют в большую сторону до ближайшего целого. Далее следует заполнить частотную таблицу (табл. 3).

Таблица 3 – Частотная таблица вариационного ряда по интервалам

№ ин-ла	**Начало** x_i	**Середина** $\tilde{x}_i$	**Конец** x_{i+1}	**Частота** m_i	**Эмпир. вер.** p_i^*
1	$x_1 = x_{\min}$	$\tilde{x}_1 = \frac{x_1 + x_2}{2}$	$x_2 = x_1 + l$	m_1	m_1 / n
2	x_2	$\tilde{x}_2 = \frac{x_2 + x_3}{2}$	$x_3 = x_2 + l$	m_2	m_2 / n
…	…	…	…	…	…
K	x_K	$\tilde{x}_K = x_K + \frac{l}{2}$	x_{max}	m_K	m_k / n
Проверка				$\sum m_i = n$	1

Затем строить соответствующие гистограммы.

Статистические оценки параметров распределения

Генеральная совокупность может характеризоваться типом распределения и некоторыми параметрами распределения (математическое ожидание, дисперсия и др.). По выборкам можно найти оценки этих параметров. Для нахождения оценок этих параметров используют *статистики* – функции от выборочных значений. Распространёнными примерами статистик являются:

выборочное среднее $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$;

выборочная дисперсия $S_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$;

выборочный *k*-тый начальный момент $\bar{m}_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k$;

выборочный *k*-тый центральный момент $\bar{\mu}_k = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^k$ и др.

Поскольку результаты опытов случайны, *любая статистика* представляет собой *случайную величину*. Чтобы статистика могла служить оценкой данного параметра θ, необходимо, чтобы распределение этой статистики было сосредоточено в достаточной близости от неизвестного значения параметра θ, т.е. так, чтобы вероятность больших отклонений этой статистики от θ была достаточно мала. Желательно также, чтобы точность оценивания увеличивалась при увеличении объема выборки.

Пусть $\hat{\theta}_n$ – некоторая статистическая оценка, полученная по выборке и оценивающая неизвестный параметр θ распределения генеральной совокупности. Если оценка определяется одним числом $\hat{\theta}_n$, то ее называют *точечной*; если вычисляют две величины, θ_{1n} и θ_{2n}, такие, что $\theta_{1n} \le \theta \le \theta_{2n}$, то такую оценку для θ называют *интервальной*.

2.1 Свойства оценок

К точечным оценкам предъявляются следующие требования:

- Оценка должна быть *несмещенной*, т.е. не иметь систематической погрешности, т.е. $M(\hat{\theta}_n) = \theta$. Соответственно, смещенной называют такую оценку, математическое ожидание которой не равно оцениваемому параметру;

- Оценка должна быть *эффективной*, т.е. такой, которая при заданном объеме выборки n имеет наименьшую возможную дисперсию. Эффективность оценок сильно зависит от распределения генеральной совокупности, если распределение нормальное, то выборочная средняя и дисперсия будут эффективными оценками;
- Оценка должна быть *состоятельной*, т.е. должно выполнятся условие: при неограниченном росте объема выборки оценка $\hat{\theta}_n$ стремится к параметру θ по вероятности, или при $n \to \infty$ для произвольного $\varepsilon > 0$ $P\left(\left|\hat{\theta}_n - \theta\right| > \varepsilon\right) \xrightarrow[n\to\infty]{} 0$.

Нужно отметить, если некоторая оценка $\hat{\theta}_n$ является состоятельной, то она асимптотически несмещенная, однако обратное утверждение не верно, т.е. условие состоятельности является более сильным, чем условие несмещенности.

2.2 Интервальное оценивание параметров распределеня

Точечные оценки имеют тот недостаток, что по ним нельзя судить о точности получаемых оценок. Поэтому возникает задача определения на основании выборочных значений такого интервала (θ_1, θ_2), который покрывал бы неизвестное значение параметра θ с заданной вероятностью.

Пусть $P(\theta_1 \le \theta \le \theta_2) = \alpha$, где случайный интервал (θ_1, θ_2), который называется *доверительным интервалом*, с заданной вероятностью α содержит оцениваемый параметр θ. Величину α называют *доверительным уровнем* или *надежностью*. Величина $\delta = (\theta_2 - \theta_1)/2$ характеризует *точность* интервальной оценки. Обычно величину α берут равной 0,95, 0,99 или 0,999. Величину $1 - \alpha$ называют *уровнем значимости* отклонения оценки. Концы доверительного интервала θ_1 и θ_2 называют *доверительными границами*.

Один из распространенных методов построения доверительных интервалов заключается в следующем. По выбранным значениям вычисляется несмещенная точечная оценка $\hat{\theta}_n$ параметра θ. Каким-либо образом вычисляется дисперсия статистики $\hat{\theta}_n$ или ее оценка $\hat{\sigma}_n^2$. Затем строят доверительный интервал вида $\left(\hat{\theta}_n - k_1\hat{\sigma}_n, \hat{\theta}_n + k_2\hat{\sigma}_n\right)$, где k_1 и k_2 – коэффициенты, значения которых определяют выбранный доверительный уровень и априорные предположения о распределении генеральной

совокупности (например, нормальность или симметричность распределения). Но поскольку такой интервал определяется не однозначно, накладываются дополнительные условия, чтобы данный интервал имел минимальную длину. Если распределение статистики $\hat{\theta}_n$ симметрично (или близко к симметричному), то доверительный интервал минимальной длины получается при $k_1 = k_2$. На такой основе строится известный критерий Стьюдента для нормально распределенных генеральных совокупностей. В самом общем случае (при минимальных предположениях относительно распределения генеральной совокупности) доверительные интервалы можно построить на основании неравенства Чебышева или других подобных неравенств. Однако такие интервальные оценки имеют небольшую точность.

2.2.1 Интервальные оценки математического ожидания

Наиболее общая статистическая модель: произвольное распределение генеральной совокупности с конечной известной дисперсией σ^2. Доверительный интервал для неизвестного математического ожидания, построенный на основании неравенства Чебышева, имеет вид $\left(\bar{x} - k\frac{\sigma}{\sqrt{n}}, \bar{x} + k\frac{\sigma}{\sqrt{n}}\right)$, где $k = 1/\sqrt{1-\alpha}$ рассчитывают в соответствии с заданным доверительным уровнем α.

Статистическая модель 2: генеральная совокупность имеет симметричное одномодальное распределение с известной конечной дисперсией σ^2. Воспользуемся неравенством Гаусса и получим следующий доверительный интервал $\left(\bar{x} - k\frac{\sigma}{\sqrt{n}}, \bar{x} + k\frac{\sigma}{\sqrt{n}}\right)$, где $k = \frac{3}{2\sqrt{1-\alpha}}$.

Статистическая модель 3: произвольное распределение генеральной совокупности с конечным четвертым моментом и неизвестной дисперсией при $n > 30$. Доверительный интервал в данной статистической модели строится на основе асимптотической нормальности оценок и имеет вид $\left(\bar{x} - k\frac{S_n}{\sqrt{n}}, \bar{x} + k\frac{S_n}{\sqrt{n}}\right)$. Величину k определяют из уравнения $\alpha = 2\Phi(k) - 1$, где α − заданный доверительный уровень, Φ − функция распределения стандартного нормального закона. Отметим, что применение вместо нормального распределения Стьюдента расширяет доверительный интервал, тем самым повышая его надежность, поэтому его используют чаще.

Статистическая модель 4: генеральная совокупность имеет нормальное распределение с математическим ожиданием μ и дисперсией σ^2. Способ построения доверительного интервала для математического ожидания зависит от того, известно ли значение дисперсии σ^2. Если это значение известно, то доверительный интервал имеет вид $\left(\bar{x} - k\frac{\sigma}{\sqrt{n}}, \bar{x} + k\frac{\sigma}{\sqrt{n}}\right)$, где k определяют из уравнения $\alpha = 2\Phi(k) - 1$. Если значение дисперсии неизвестно, то вместо него используют выборочную дисперсию $S_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$, а k определяют из уравнения $\alpha = 2F_{n-1}(k) - 1$, где F_{n-1} – функция распределения Стьюдента с $(n-1)$ степенями свободы. Доверительный интервал имеет вид $\left(\bar{x} - k\frac{S_n}{\sqrt{n-1}}, \bar{x} + k\frac{S_n}{\sqrt{n-1}}\right)$.

2.2.2 Интервальные оценки дисперсии

Способы получения интервальных оценок дисперсии сильно зависят от типа распределения генеральной совокупности, поэтому приведем наиболее общие из них.

Статистическая модель 1: произвольное распределение генеральной совокупности с конечным четвертым моментом. Объем выборки не менее 50. Поскольку нет априорных предположений о типе распределения генеральной совокупности, следует использовать асимптотическую нормальность распределения статистик для вычисления моментов генеральной совокупности. В этом случае доверительный интервал имеет вид $\left(S_n^2 - k\sigma\left(S_n^2\right), S_n^2 + k\sigma\left(S_n^2\right)\right)$, где коэффициент k определяют из уравнения $\alpha = 2\Phi(k) - 1$, где α − заданный доверительный уровень, Φ − функция распределения стандартного нормального закона. Среднеквадратическое отклонение $\sigma\left(S_n^2\right)$ статистики S_n^2 вычисляется по формуле $\sigma\left(S_n^2\right) = \sqrt{\frac{\bar{\mu}_4 - S_n^2}{n}}$, где $\bar{\mu}_4 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^4$.

Статистическая модель 2: генеральная совокупность имеет нормальное распределение с математическим ожиданием μ и дисперсией σ^2. В этом случае доверительный интервал имеет вид $\left(\frac{n}{t_{\hat{A}}}S_n^2, \frac{n}{t_{\acute{I}}}S_n^2\right)$, где $t_B = F_{n-1}^{-1}(\beta_B)$,

$t_H = F_{n-1}^{-1}(\beta_H)$, в свою очередь $\beta_B = (1+\alpha)/2$, $\beta_H = (1-\alpha)/2$, а F_{n-1}^{-1} – функция обратная к функции распределения χ^2 с $(n-1)$ степенями свободы. Если значение математического ожидания не известно, то вычисляются следующие точечные оценки $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, а затем $S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$. Если математическое ожидание известно и равно μ, то $S_n^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \mu(2\bar{x} - \mu)$.

2.3 Точечное оценивание параметров распределения генеральной совокупности по вборочным данным

Существует несколько подходов к оценке параметров распределения генеральной совокупности. Мы остановимся на двух из них.

2.3.1 Точечное оценивание генеральной средней и генеральной дисперсии непосредственно по выборке

Генеральной средней $\bar{x}_Г$ называют среднее арифметическое значений признака генеральной совокупности. Если все значения x_1, x_2, …, x_N признака генеральной совокупности объема N ***различны***, то $\bar{x}_Г = \frac{x_1 + x_2 + ... + x_N}{N}$. Если же значения признака x_1, x_2, …, x_k имеют соответствующие частоты N_1, N_2, …, N_k, причем $N_1 + N_2 + ... + N_k = N$, то $\bar{x}_Г = \frac{x_1 N_1 + x_2 N_2 + ... + x_k N_k}{N}$, т.е. генеральная средняя есть средняя взвешенная значений признака с весами, равными соответствующим частотам.

Выборочной средней $\bar{x}_в$ называют среднее арифметическое значений признака выборочной совокупности. Если все значения x_1, x_2, …, x_n признака выборочной совокупности объема n ***различны***, то $\bar{x}_в = \frac{x_1 + x_2 + ... + x_n}{n}$. Если же значения признака x_1, x_2, …, x_k имеют соответствующие частоты m_1, m_2, …, m_k, причем $m_1 + m_2 + ... + m_k = n$, то $\bar{x}_в = \frac{x_1 m_1 + x_2 m_2 + ... + x_k m_k}{n}$.

Выборочная средняя является несмещенной оценкой генеральной средней.

Генеральной дисперсией $D_Г$ называют среднее арифметическое квадратов отклонений значений признака генеральной совокупности от их среднего значения $\bar{x}_Г$.

Выборочной дисперсией $D_в$ называют среднее арифметическое квадратов отклонений значений признака выборочной совокупности от их среднего значения $\bar{x}_в$. Найти выборочную дисперсию можно по формуле $D_в = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_в)^2}{n}$ или по формуле $D_в = \frac{\sum_{k=1}^{n} x_i^2}{n} - \bar{x}_в^2$.

Выборочная дисперсия является смещенной оценкой генеральной дисперсии, поскольку занижает значение генеральной дисперсии, особенно при незначительном *n*

$$M(D_в) = \frac{n-1}{n} D_Г .$$

«Исправленная дисперсия» будет иметь вид $D_Г = \bar{S}^2 = \frac{n}{n-1} D_в$.

Естественно, что генеральное среднеквадратическое отклонение (стандарт) вычисляется $\bar{S} = \sqrt{D_Г}$.

2.3.2 Точечное оценивание выборочной средней и выборочной дисперсии по данным частотной таблицы

Допустим мы сформировали следующий вариационный ряд:

№ ин-ла	**Начало** x_i	**Середина** $\tilde{x}_i$	**Конец** x_{i+1}	**Частота** m_i	**Эмпир. вер.** p_i^*
1	$x_1 = x_{\min}$	$\tilde{x}_1 = \frac{x_1 + x_2}{2}$	$x_2 = x_1 + l$	m_1	m_1 / n
2	x_2	$\tilde{x}_2 = \frac{x_2 + x_3}{2}$	$x_3 = x_2 + l$	m_2	m_2 / n
…	…	…	…	…	…
K	x_K	$\tilde{x}_K = x_K + \frac{l}{2}$	x_{max}	m_K	m_k / n
Проверка				$\sum m_i = n$	1

Выборочное среднее можно оценить по формуле $\bar{x}_в = \frac{1}{n}\sum_{i=1}^{K} m_i \tilde{x}_i$.

Выборочную дисперсию можно оценить по формуле $D_в = \frac{1}{n}\sum_{i=1}^{K} m_i \tilde{x}_i^2 - \bar{x}_в^2$.

3 Статистическая проверка статистических гипотез

Статистической называют гипотезу о виде неизвестного распределения или о параметрах известного распределения. Например,

- генеральная совокупность распределена по нормальному закону распределения;
- математическое ожидание нормально распределенной генеральной совокупности равно 10;
- дисперсии двух нормальных совокупностей равны между собой.

Как правило, выдвигается две гипотезы H_0 – основная гипотеза, например, генеральная совокупность распределена по равномерному закону распределения, и H_1 – конкурирующая (альтернативная) гипотеза, например, генеральная совокупность не распределена по равномерному закону распределения.

Гипотезы бывают *простыми* ($\lambda = 5$) или *сложными* ($\lambda > 5$).

Выдвинутая гипотеза может быть справедливой или ошибочной, поэтому могут возникнуть ошибки двух типов:

Ошибка первого рода – отвергнута правильная гипотеза.

Ошибка второго рода – принята неправильная гипотеза.

Последствия ошибок могут быть различными. Например, если отвергнуто правильное решение «продолжать строительство жилого дома», то эта ошибка *первого* рода повлечет *материальный ущерб*; если же приняли неправильное решение «продолжать строительство», то эта ошибка *второго* рода может повлечь *гибель людей*. Вероятность совершить ошибку первого рода обозначают α, а вероятность ошибки второго рода – β. Любую статистическую гипотезу необходимо статистически проверить.

3.1 Статистические гипотезы и их проверка

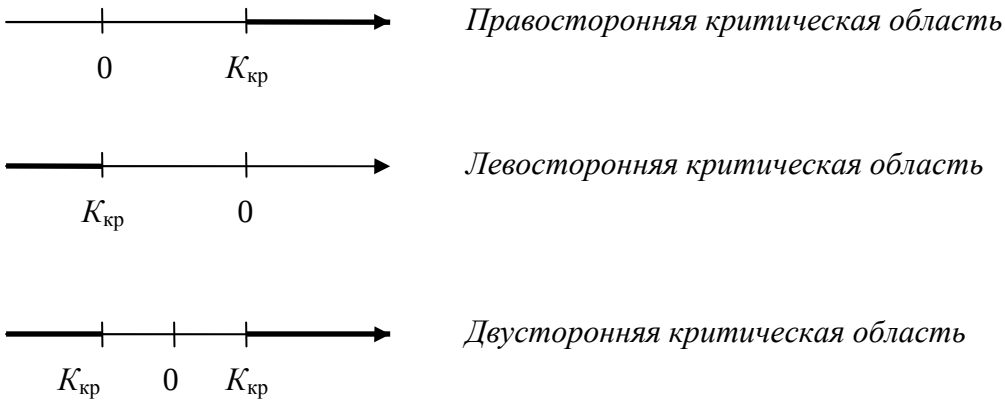
Статистические гипотезы проверяются с помощью статистических критериев. Для проверки основной гипотезы используют специально подобранную случайную величину *K* (статистику), точное или приближенное значение которой известно. ***Критерием согласия*** называют критерий проверки гипотезы о принадлежности распределения генеральной совокупности конкретному классу распределений. Имеется несколько критериев согласия: χ^2 («хи квадрат») К. Пирсона, критерий Колмогорова (применяется только для непрерывных распределений), критерий Смирнова.

После выбора определенного критерия, множество всех его возможных значений разбивают на два непересекающихся подмножества: одно из них содержит значения критерия, при которых основная гипотеза отвергается (*критическая область*), а другое – при котором она принимается (*область принятия гипотезы*).

Основной принцип проверки статистических гипотез: если $K_{набл}$ принадлежит критической области – основную гипотезу отвергают, в противном случае основную гипотезу принимают.

$K_{кр}$ – точка, отделяющая критическую область от области принятия гипотезы.

Различают *одностороннюю* (правостороннюю, левостороннюю) и *двустороннюю* критическую область.



Для правосторонней критической области точку $K_{кр}$ ищут исходя из требования, чтобы при условии справедливости H_0, вероятность того, что критерий $K_{набл}$ примет значение, большее $K_{кр}$, была равна уровню значимости α.

3.2 *Критерий согласия Пирсона*

Достоинством критерия Пирсона является его универсальность – применяется одинаково при разных распределениях.

К Пирсоном доказано, что при $n \to \infty$ закон распределения случайной величины $\chi^2 = \sum \frac{\left(m_i - n_i^T\right)^2}{n_i^T}$, не зависимо от того, какому закону распределения подчинена генеральная совокупность, стремится к закону

распределения $\chi^2 = \sum_{i=1}^{n} X_i^2$ с $k=n$ степенями свободы, где X_i - нормальные независимые случайные величины, причем математическое ожидание каждой из них равно нулю, а среднее квадратичное отклонение – единице. Дифференциальная функция распределения χ^2 имеет следующий вид:

$$f(x) = \begin{cases} 0, & x \le 0 \\ \dfrac{1}{2^{\frac{k}{2}} \cdot \Gamma\left(\frac{k}{2}\right)} e^{-\frac{x}{2}} \cdot x^{\frac{k}{2}-1}, & x > 0 \end{cases}, \text{ где } \Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt \text{ – гамма-функция,}$$

$$\Gamma(n+1)=n!$$

С увеличением числа степеней свободы распределение χ^2 медленно приближается к нормальному.

Идея критерия Пирсона состоит в следующем: наблюдаемые (эмпирические) частоты m_i сравниваются с теоретическими n'_i (вычисленные в предположении о проверяемом законе распределения). Обычно эмпирические и теоретические частоты различаются. Возможно, что расхождение частот *случайно* (незначимо) и объясняется малым числом наблюдений, либо способом их группировки, или другими причинами. Возможно, расхождение частот неслучайно (*значимо*) и объясняется тем, что теоретические частоты вычислены исходя из *неверной* гипотезы. Критерий Пирсона устанавливает, на принятом уровне значимости α (достаточно малая вероятность ошибки первого рода), *согласуются* или *не согласуются* данные наблюдений с выдвинутой гипотезой.

Критерий Пирсона имеет правостороннюю критическую область. Вероятность попадания критерия χ^2 в критическую область, в предположении справедливости нулевой гипотезы, равна принятому уровню значимости α:

$$P\left(\chi^2 > \chi^2_{кр}(\alpha, k)\right) = \alpha .$$

Таким образом, правосторонняя критическая область определяется неравенством $\chi^2 > \chi^2_{кр}(\alpha, k)$, а область принятия нулевой гипотезы – неравенством $\chi^2 \le \chi^2_{кр}(\alpha, k)$.

Правило. Для того чтобы при заданном уровне значимости, проверить нулевую гипотезу H_0: *генеральная совокупность распределена по закону A*, надо:

1. В предположении, что справедлива гипотеза H_0, вычислить теоретические частоты n_i';
2. Вычислить наблюдаемое значение критерия $\chi^2_{набл} = \sum \frac{(m_i - n'_i)^2}{n'_i}$;
3. По заданному уровню значимости α, и числу степеней свободы k ($k = s - 1 - r$, где s – число групп или частичных интервалов выборки, r – число параметров предполагаемого распределения) по таблице распределения критических точек распределения χ^2 найти критическую точку $\chi^2_{кр}(\alpha, k)$;
4. Если $\chi^2_{набл} \le \chi^2_{кр}$ – нет оснований отвергнуть нулевую гипотезу. Если $\chi^2_{набл} > \chi^2_{кр}$ – нулевую гипотезу отвергают.

Замечание 1. Объем выборки должен быть достаточно велик, не менее 50. Каждая группа должна содержать не менее 5-8 вариант, малочисленные группы следует объединить в одну, суммируя частоты.

Замечание 2. Поскольку возможны ошибки первого и второго рода, в особенности, если согласование теоретических и эмпирических частот «слишком хорошее», следует проявлять осторожность. Например, можно повторить опыт, увеличить число наблюдений, воспользоваться другими критериями, построить график распределения, вычислить асимметрию и эксцесс.

Замечание 3. В целях контроля вычислений, наблюдаемое значение критерия вычисляют по $\chi^2_{набл} = \sum \frac{m_i^2}{n_i^T} - n$.

4 Основы корреляционно-регрессионого анализа

Во время статистических наблюдений для каждого объекта в ряде случаев можно измерить значения нескольких признаков. Таким образом, получается *многомерная выборка*. Если многомерную выборку обработать по значениям отдельного признака, то получится обычная обработка одномерной выборки.

Смысл обработки многомерной выборки состоит в том, чтобы *установить связи между признаками*. Связи могут быть *функциональными* (заданные функцией), т.е. каждому значению одной величины соответствует определенное значение другой величины. Связь называется *стохастической* (или *статистической*), если изменение одной величины вызывает изменение *распределения* другой величины. Если среднее значение одной случайной величины функционально зависит от значений другой случайной величины, то такая статистическая зависимость называется *корреляционной*.

Корреляционный анализ заключается в установлении *вида фактической зависимости* случайных величин по результатам их измерений. Конкретный вид функциональной зависимости между величинами X и Y называют *эмпирической формулой*. Простейшая эмпирическая формула это линейная функция $y = ax + b$, более сложная – $y = ax^2 + bx + c$ и т.д. Задача получения эмпирической формулы состоит в нахождении коэффициентов a, b, c и т.д. Для нахождения коэффициентов используются: *метод натянутой нити*, *метод сумм*, *метод наименьших квадратов*.

Для установления зависимости и тесноты связи между случайными величинами X и Y используют *корреляционные таблицы*.

Если в двумерной случайной величине зафиксировать значение одной случайной величины, например $Y = y$, то совокупность соответствующих значений другой случайной величины X можно рассматривать как отдельную случайную величину $E(X \mid Y = y)$ со своим законом распределения и своими числовыми характеристиками, которые, как и само распределение, называют *условными*. Если рассмотреть условные средние значения одной случайной величины при всех значениях другой случайной величины, то получим следующие функции $f(x) = E(Y \mid X = x)$ и $g(y) = E(X \mid Y = y)$, называемые *функциями регрессии*. Если функции регрессии известны, то можно по значению одной случайной величины *прогнозировать* значения другой случайной величины. Обычно конкретный вид функции регрессии неизвестен и определяется по двумерной выборке. Если функция регрессии линейна $y = f(x) = \alpha_1 x + \alpha_0$, $x = g(y) = \beta_1 y + \beta_0$, то коэффициенты α_1, α_0, β_1, β_0 целесообразно находить с помощью *метода наименьших квадратов*.

Свойства коэффициента корреляции

1. Если случайные величины X и Y независимы, то теоретически коэффициент корреляции равен нулю. Противоположное утверждение неверно и не всегда выполняется.
2. Коэффициент корреляции одинаков и в случае, когда X зависит от Y, и в случае, когда Y зависит от X.
3. Если случайные величины X и Y линейно зависимы, т.е. $Y = aX + b$, то $r = \begin{cases} +1, a > 0 \\ -1, a < 0 \end{cases}$.
4. Если значение $|r|$ близко к единице, то надо найти линейную функцию регрессии. Если $|r| < 0.5$, то надо искать нелинейную функцию регрессии.

5 Основы дисперсионного анализа

Дисперсионный анализ − это статистический метод анализа результатов наблюдений, зависящих от различных факторов, *определение наиболее влияющих факторов* и *оценка* (значимость) этого *влияния*. Факторами обычно называют внешние условия (причины), влияющие на результаты наблюдений.

Дисперсионный анализ заключается в разложении общей вариации (дисперсии) наблюдаемой случайной величины на отдельные слагаемые, каждое из которых характеризует влияние того или иного фактора. Если исследуется влияние одного фактора, то говорят об однофакторном дисперсионном анализе, при исследовании двух факторов – о двухфакторном анализе. Например, несколько станков в цехе выполняю одинаковые операции. Для планирования дальнейшей обработки деталей нужно знать, все ли станки дают одинаковую продукцию, или нет. Можно ли игнорировать влияние фактора (т.е. станков) на продукцию или нет?

Статистическая модель дисперсионного анализа состоит в следующем. Наблюдаются n случайных величин $X_1, X_2, ..., X_n$, каждая из которых представима в виде $X_i = \mu + \beta_1 + \beta_2 + ... + \beta_m + \varepsilon_i$, где μ − константа (общее среднее), β_j − значение j-го фактора, ε_i − «остаточная» случайная величина, представляющая ошибки наблюдений, влияние неучтенных факторов и т.п. Как правило, предполагается, что случайные величины ε_i независимы между собой, одинаково распределены по нормальному закону с нулевым математическим ожиданием. Факторы обычно являются классификационными или порядковыми

(не количественными) величинами, принимающими конечное множество значений. В таком случае, когда β_j принимает конкретное k-е значение из этого множества, говорят о k-м уровне j-го фактора.

Цель дисперсионного анализа заключается в оценке адекватности модели имеющимся выборочным значениям (для чего определяются статистические характеристики случайных величин ε_i), а также в оценке влияния факторов (проверяются гипотезы о равенстве математических ожиданий случайных величин $X_1, X_2, ..., X_n$). Модель, в которой все β_j являются детерминированными, называется моделью с постоянными факторами. Если все β_j − случайные величины, модель со случайными факторами. Модель может быть смешанной.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. *Гмурман В.Е.* Теория вероятностей и математическая статистика, М., «Высшая школа», 2002, 1972.

2. *Красс М.С., Чупрынов Б.П.* Математические методы и модели для магистрантов экономики: Учебное пособие. – СПб.: Питер, 2006. – 496 с. ил.

3. *Минько А.А.* Статистический анализ в MS EXCEL. – М.: Издательский дом «Вильямс», 2004. – 448 с. ил.

Навчальне видання

БІЛОГУРОВА Ганна Вікторівна

ТЕОРІЯ ЙМОВІРНОСТЕЙ ТА МАТЕМАТИЧНА СТАТИСТИКА

ЧАСТИНА 2. МАТЕМАТИЧНА СТАТИСТИКА

КОНСПЕКТ ЛЕКЦІЙ

(для студентів 2-го курсу денної та заочної форм навчання, напрямів підготовки 6.030504 – Економіка підприємства, 6.030509 – Облік і аудит)

(рос. мовою)

Відповідальний за випуск *М. І. Самойленко*

За авторською редакцією

Комп'ютерне верстання *І. В. Волосожарова*

План 2011, поз. 183Л

Підп. до друку 09.12.2011	Формат 60х84/16
Друк на ризографі	Ум. друк. арк. 1,5
Тираж 50 пр.	Зам. №

Видавець і виготовлювач:
Харківський національний університет
міського господарства імені О. М. Бекетова,
вул. Революції, 12, Харків, 61002
Електронна адреса: rectorat@kname.edu.ua
Свідоцтво суб'єкта видавничої справи:
ДК № 4705 від 28.03.2014 р.